Strojno prevođenje u odabranoj domeni

Ravkin, Iris

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2020./2021.

Iris Ravkin

Strojno prevođenje u odabranoj domeni

Završni rad

Mentor: prof. dr. sc. Sanja Seljan

Zagreb, rujan 2021.

Izjava o akademskoj čestitosti

Izjavljujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

potpis

_____

Iris Ravkin

Sadržaj

1. Uvod

Razvoj strojnog prevođenja vezan je uz razvoj tehnologije, ali i potreba koje se odnose na brzu komunikaciju i prijenos informacija. Pri tome tehnologija automatskog strojnog prevođenja i strojno potpomognutog prevođenja ima veliku ulogu u različitim područjima primjene, kao što je poslovanje (Seljan, 2011[1]), prijenos informacija s jednog jezika na drugi (Seljan i Dunđer, 2014[2]), dostup do informacija preko weba i različitih aplikacija, kroz npr. lokalizaciju igara (Seljan i Katalinić, 2017[3]) , učenje jezika (Kučiš i Seljan, 2014[4]; Seljan, 2019[5]) te u kombinaciji s ostalim tehnologijama, kao što je npr. sažimanje teksta (Seljan i sur., 2015[6]) ili glasovne tehnologije (Dunđer i sur., 2013[7]).

Prevođenje je kompleksni proces koji uključuje puno više od prenošenja riječi iz jednog jezika u drugi. Ono zahtijeva znanje o vokabularu izvornog i ciljnog jezika, kao i znanje o sustavima pravila prema kojima se formiraju rečenice, ističu Arnold i sur. (1994)[8]. Osim toga, potrebno je i znanje o svijetu, znanje o prirodi stvari i načelima koje one slijede. Također je potrebno i tehničko znanje o temi u pitanju, te o domeni teksta. Kako bi se osigurala kvaliteta prevedenog teksta, provodi se evaluacija prijevoda. Evaluacija je proces koji se može provoditi primjenom različitih metrika, kao što je evaluacija sustava za strojno potpomognuto prevođenje (Seljan i sur., 2021[9]), a kompanije koje se bave ovom tehnologijom imaju za cilj provoditi upravljanje

[1] Seljan, S. (2011). Translation technology as Challenge in education and business. *Informatologia*, *44*(4), 279–286.
[2] Seljan, S., Dunđer, I. (2014). Combined Automatic Speech Recognition and Machine Translation in Business Correspondence Domain for English-Croatian. *International Journal of Computer and Information Engineering, World Academy of Science, Engineering and Technology, 8*(11)*,* 1980-1986.
[3] Seljan, S., Katalinić, J. (2017). Integrating Localization into a Video Game. INFuture2017: Integrating ICT in Society.
[4] Kučiš, V., Seljan, S. (2014). The role of online translation tools in language education. *Babel*, *60*(3).
[5] Seljan, S. (2019) Informacijska i komunikacijska tehnologija (IKT) u interdisciplinarnom okruženju nastave jezika. U: Vrhovac, Y., Berlengi Kapučin, V., Geld, R., Jelić, A., Letica Krevelj, S., Mardečić, S., Lütze-Miculinić, M. (ur.) *Izazovi učenja stranoga jezika u osnovnoj školi* (pp. 446–461). Ljevak.
[6] Seljan, S., Klasnić, K., Stojanac, M., Pešorda, B., Mikelić Preradović, N. (2015). Information Transfer through Online Summarizating and Translation Technology. INFuture2015: e-Institutions–Openness, Accessibility, and Preservation.
[7] Dunđer, I., Seljan, S., Arambašić, M. (2013). Domain-specific Evaluation of Croatian Speech Synthesis in CALL. Recent Advances in Information Science - Computer Engineering, WSEAS 1, 142-147.
[8] Arnold, D., Balkan, L., Meijer, S., Lee Humphreys, R., Sadler, L. (1994). Machine translation: an introductory guide. Blackwells NCC.
[9] Seljan, S., Škof Erdelja, N., Kučiš, V., Dunđer, I., Pejić Bach, M. (2020). Quality Assurance in Computer-Assisted Translation in Business Environments. Natural Language Processing for Global and Local Business. IGI-Global, 242-270.

kompletnim sustavom tijeka prevođenja u manjim kompanijama i u većim institucijama (Seljan, 2018a[10]; Seljan, 2018b[11]).

No postavlja se pitanje može li računalo koje ne posjeduje znanje o svijetu generirati kvalitetne prijevode poput čovjeka, i na koji način je to moguće postići? Kroz povijest su postojali mnogi pokušaji izgradnje sustava za strojno prevođenje, ali činjenica je da oni još uvijek nisu uspjeli dostići razinu ljudskih prevoditelja, iako današnji sustavi pokazuju izuzetne rezultate. Osim navedenoga, postoje i situacije kada je dovoljan grubi prijevod, tzv. „gisting" za prenošenje osnovnog značenja, pri čemu je dovoljno korištenje automatskih sustava, a potreba za izmjenama dobivenog prevedenog teksta sve je manja. Uz razvoj tehnologije i otkrivanjem novih metoda, suvremeni sustavi za strojno prevođenje postižu sve bolje rezultate u kvaliteti prijevoda.

U teorijskom dijelu ovog rada analizirat će se tri osnovna pristupa sustavima za strojno prevođenje: sustav temeljen na pravilima, statistički sustav i neuronski sustav za strojno prevođenje. Prikazat će se glavne karakteristike i načela rada tih sustava.

U praktičnom dijelu rada prikazat će se proces izgradnje vlastitog sustava za strojno prevođenje iz određene domene, korištenjem odabranog alata. Na kraju će se prikazati rezultati i vrednovati kvaliteta dobivenih prijevoda za hrvatsko-engleski jezični par. Nakon toga slijedi zaključak, popis literature, te popis slika, grafova i tablica.

[10] Seljan, S. (2018a). Total Quality Management Practice in Croatian Language Service Provider Companies. EntreNova 18, 4 (1), 461-469. INFuture2015: eInstitutions–Openness, Accessibility, and Preservation.
[11] Seljan, S. (2018b). Quality Assurance (QA) of Terminology in a Translation Quality Management System (QMS) in the Business Environment. European Parliament: Translation Services in the Digital World, 92-145.

2. O sustavima za strojno prevođenje

Poibeau (2017)[12] navodi kako su se brojni znanstvenici kroz povijest bavili pitanjem jezične raznolikosti. Nastojali su premostiti tu raznolikost na mnoge načine, ali jedno od najznačajnijih otkrića za unaprjeđenje prevođenja bio je razvoj računala. Nakon Drugog svjetskog rata, razvili su se prvi sustavi za strojno prevođenje, tj. sustavi sposobni za automatski prijevod teksta s izvornog jezika na ciljni jezik.

Prema Dovedan i sur. (2002)[13], prvi uređaji za strojno prevođenje stvoreni su početkom tridesetih godina kroz patente koje su neovisno iznijeli George Artsrouni i Petr Smirnov-Troyanskii. Ozbiljniji pokušaji započeli su nakon pojave ENIACa (engl. *Electronic Numerical Integrator and Calculator*) 1946. godine. Tada nastaju mnoge teorije vezane za samo područje strojnog prevođenja, ali i za informacijske znanosti općenito.

Jedan od začetnika teorije komunikacije, Warren Weaver iznio je teoriju da bi se tehnike vojnog šifriranja mogle primijeniti i na strojno prevođenje. Dakle, računalo je trebalo svaku riječ pronaći u dvojezičnom rječniku, usporediti riječi iz ulaznog teksta s onima pohranjenima u rječniku (binarno ili slučajno pretraživanje), odabrati odgovarajući prijevod i nakon obrade cijele rečenice, prevedene riječi složiti u određeni poredak prema pravilima ciljnoga jezika. Za razliku od programa koji samo "gledaju" riječi, sustav za strojno prevođenje treba raščlaniti tekst izvornog jezika i slagati rečenice u ciljnome jeziku (Dovedan i sur., 2002[14]).

Proučavanje jezika usko je povezano s analizom znanja i rasuđivanja, a upravo zbog toga su interes za ovo polje iskazali mnogi znanstvenici iz drugih disciplina poput filozofije, lingvistike, računalnih znanosti, te umjetne inteligencije. Njihov je utjecaj vidljiv kroz cijelu povijest razvoja sustava za strojno prevođenje.

Prema Koehnu (2020)[15], sustavi za strojno prevođenje nastoje postići kvalitetne prijevode, a kvaliteta prijevoda mjeri se pomoću metrika adekvatnosti i fluentnosti. Adekvatnost označava očuvanje značenja originalnog teksta, dok fluentnost označava produkciju teksta koji izgleda poput originalnog teksta na ciljnom jeziku. Određene domene tekstova, poput književnih djela

[12] Poibeau, T. (2017). *Machine translation*. The MIT Press.
[13] Dovedan, Z., Seljan, S., Vučković, K. (2002). Strojno prevođenje kao pomoć u procesu komunikacije. *Informatologia*, 35(4).
[14] Ibid.
[15] Koehn, P. (2020). *Neural machine translation*. Cambridge University Press.

ili poezije, pridaju veću važnost stilu i tečnosti teksta, dok se doslovno značenje nekih rečenica može i promijeniti bez da se pritom gubi cjelokupan dojam teksta. S druge strane, kod prevođenja drugih vrsta tekstova, poput stručnih, pravnih tekstova ili tehničkih uputa, prvenstveno je bitno značenje i točnost činjenica izraženih u tekstu, čak i ako u tom slučaju tekst može izgubiti na tečnosti. Zbog toga je kod prevođenja bitno obratiti pozornost na domenu teksta.

Različiti sustavi za strojno prevođenje nastoje postići prijevode što sličnije ljudskima, iako potpuno automatizirani sustavi još uvijek nisu na toj razini. Ali razvojem novih pristupa strojnom prevođenju, razlika između strojnog i ljudskog prijevoda postaje manja, a kvaliteta prijevoda bolja, pri čemu je jedna od metoda ograničavanje na određenu domenu u kojoj se koristi određena vrsta podjezika (Seljan, 2000[16]; Seljan i Dunđer, 2015a[17]).

Evaluacija strojnih prijevoda može biti ljudska ili automatska. Prema Seljan i sur. (2015)[18], ljudska evaluacija ocjenjuje prijevode prema kriterijima fluentnosti i adekvatnosti. Takav način evaluacije subjektivan je i predstavlja naporan zadatak koji zahtjeva puno vremena i novca. S druge strane, automatska evaluacija provodi se pomoću raznih metrika. Njima se vrednuje kvaliteta prijevoda usporedbom strojnog prijevoda i referentnog prijevoda istog teksta.

Seljan i Dunđer (2015a)[19] također navode i glavne prednosti metrika za automatsku evaluaciju, a to su brzina, cijena i objektivnost. Metrike uvijek rade na isti način, mogu se podešavati i pružaju smislene, dosljedne, točne i pouzdane informacije o razini kvalitete strojnog prijevoda.

Neke od najpoznatijih automatskih metrika za evaluaciju strojnih prijevoda su BLEU, NIST, F-measure (GTM) i METEOR. BLEU (BiLingual Evaluation Understudy) je najčešće korištena metrika za automatsku evaluaciju prijevoda (Seljan i Dunđer, 2015b[20]).

[16] Seljan, S. (2000). Sublanguage in Machine Translation. Mipro 2000: Computers in Intelligent Systems
[17] Seljan, S., Dunđer, I. (2015a). Automatic quality evaluation of machine-translated output in sociological-philosophical-spiritual domain. Central European Conference on Information and Intelligent Systems (CECIIS).
[18] Seljan, S., Tucaković, M., Dunđer, I. (2015). Human Evaluation of Online Machine Translation Services for English/Russian-Croatian. *New Contributions in Information Systems and Technologies.*
[19] Seljan, S., Dunđer, I. (2015a). Automatic quality evaluation of machine-translated output in sociological-philosophical-spiritual domain. Central European Conference on Information and Intelligent Systems (CECIIS).
[20] Seljan, S., Dunđer, I. (2015b). Machine Translation and Automatic Evaluation of English/Russian-Croatian. Proceedings of the International Conference "Corpus Linguistics". St. Petersburg State University, 72-79.

Sustavi za strojno prevođenje suočavaju se s mnogim problemima. Uz mnoge sintaktičke i semantičke prepreke, Koehn (2020)[21] navodi da je najveći problem kod obrade prirodnog jezika, pa tako i kod strojnog prevođenja, višeznačnost. Prirodni jezici višeznačni su na svakoj razini: značenju riječi, morfologiji, sintaktičkim svojstvima i ulogama, te vezama među različitim dijelovima teksta. Ljudski prevoditelji mogu lakše razriješiti problem višeznačnosti budući da su upoznati s kontekstom i imaju šire znanje o temi, za razliku od računala.

[21] Koehn, P. (2020). *Neural machine translation*. Cambridge University Press.

3. Sustavi za strojno prevođenje temeljeni na pravilima

Prema Arnold i sur. (1994)[22], najraniji pristupi strojnom prevođenju vodili su se očekivanjem da bi većinu procesa uključenih u prevođenje mogao obaviti sustav koji sadrži informacije iz dvojezičnog rječnika. Riječi na izvornom jeziku bile bi zamijenjene njihovim ekvivalentima na ciljnom jeziku, na temelju ugrađenog rječnika. Također, riječi u rečenici na ciljnom jeziku bile bi organizirane prema poretku karakterističnom za ciljni jezik, na temelju određenih pravila. Dakle, bila bi potrebna samo dva koraka kako bi se dobio točan prijevod spreman za upotrebu. Ovakvo viđenje procesa prevođenja slijedi pretpostavku da se prijevod sastoji samo od zamjene riječi njihovim ekvivalentima na drugom jeziku i promjene poretka riječi unutar rečenica prema pravilima. Međutim, pokazalo se da su kvalitetni sustavi za prevođenje puno kompleksniji od toga.

Kod sustava temeljenih na pravilima, pri procesu prevođenja primjenjuju se pravila koja su ustanovili ljudski stručnjaci. Bhattacharyya (2015)[23] ističe kako su sva ta pravila ograničena znanjem i stručnošću ljudi, njihovim razumijevanjem svojstava jezika kojima se bave, ali i domenom samog teksta. Sustavi temeljeni na pravilima imaju visoku preciznost i nizak odziv, što znači da kada su primjenjivi, većinom su i precizni, tj. daju točan prijevod. Ali s druge strane, ne postoji puno situacija u kojima su primjenjivi, što znači da imaju nizak odziv. Također, postoji mogućnost konflikta među pravilima, tj. u određenim situacijama može biti primjenjivo više pravila. Zbog toga je bitno da ta pravila budu pažljivo osmišljena. Sreelekha i sur. (2017)[24] dodaju da sustavi temeljeni na pravilima zahtijevaju velike ljudske napore kako bi se odredila pravila i pripremili resursi za obradu teksta.

Možemo zaključiti da su upravo ljudski stručnjaci zaslužni za sposobnosti i kvalitetu sustava za prevođenje temeljenih na pravilima.

[22] Arnold, D., Balkan, L., Meijer, S., Lee Humphreys, R., Sadler, L. (1994). *Machine translation: an introductory guide*. Blackwells NCC.
[23] Bhattacharyya, P. (2015). *Machine translation*. Chapman & Hall/CRC.
[24] Sreelekha, S., Bhattacharyya, P., Malathi, D. (2017). Statistical vs. Rule-Based Machine Translation: A Comparative Study on Indian Languages. *Advances in Intelligent Systems and Computing*, 663–675.

3.1. Pristupi sustavima temeljenim na pravilima

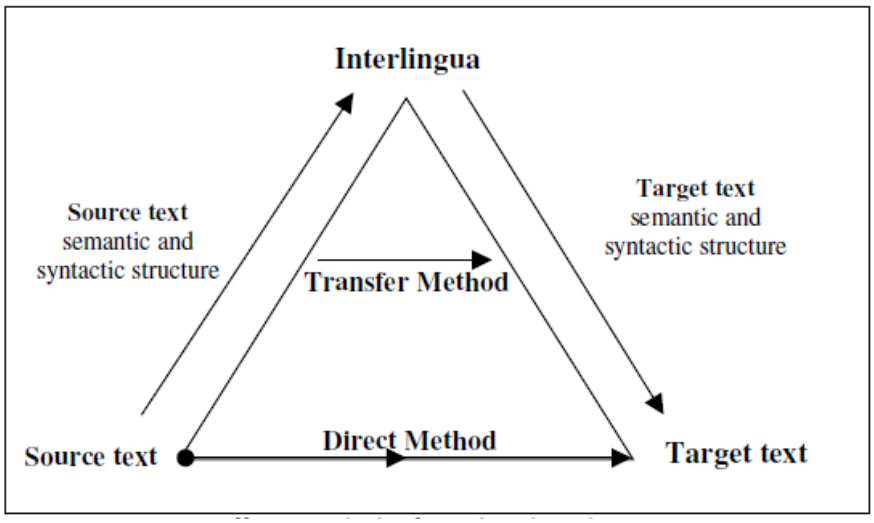
Prema Hutchinsu (1993)[25], postoje dva glavna pristupa sustavima temeljenim na pravilima: interlingua model i metoda transfera.

Interlingua model temelji se na analizi u jezično neutralni prikaz i generiranju prijevoda na temelju tog prikaza.

Metoda transfera uključuje tri faze:

1. Analiza (engl. *analysis*) u apstraktni prikaz izvornog jezika
2. Transfer (engl. *transfer*) u apstraktni prikaz ciljnog jezika
3. Generiranje (engl. *generation*) ili sinteza teksta na ciljnom jeziku

Arnold i sur. (1994)[26] nadovezuju se na tu ideju. Većina sustava temeljenih na metodi transfera ili na interlingua modelu temelje se na načelu da je za uspješni strojni prijevod potrebno definirati razinu reprezentacije teksta. Ona treba biti dovoljno apstraktna kako bi prijevod bio što izravniji, ali i dovoljno površna kako bi se rečenice u izvornim i ciljnim jezicima mogle uspješno mapirati u tu razinu reprezentacije. Dakle potreban je kompromis između dubine analize, tj. razumijevanja izvornog teksta i definiranja njegove apstraktne reprezentacije.



Slika 1. Metode sustava temeljenih na pravilima (Tripathi i Sarkhel, 2011)[27]

[25] Hutchins, W.J. (1993). *Latest Developments in Machine Translation Technology: Beginning a New Era in MT Research.*
[26] Arnold, D., Balkan, L., Meijer, S., Lee Humphreys, R., Sadler, L. (1994). *Machine translation: an introductory guide*. Blackwells NCC.
[27] Tripathi, S. i Sarkhel, J. (2011). Approaches to machine translation. *Annals of Library and Information Studies, 57.* 388-393.

3.2. Interlingua model

Poibeau (2017)[28] dalje navodi kako se pravila transfera moraju prilagoditi svakom novom jezičnom paru koji se prevodi. Ideja interlingua modela nastoji riješiti taj problem tako što dodaje razinu međujezika, tj. reprezentacije neovisne o jeziku. Kao i metoda transfera, interlingua model se također sastoji od koraka analize izvornog jezika. Ali budući da ne postoji korak transfera, u idućem koraku analiza rezultira pretvaranjem riječi izvornog jezika u međujezik. Na temelju te reprezentacije generira se prijevod u ciljni jezik.

Bhattacharyya (2015)[29] definira međujezik kao prikaz značenja teksta koje nije ovisno o jeziku, tj. prikaz koji predstavlja značenje nekog teksta bez višeznačnosti. To znači da se izvorna rečenica mora svesti na samo jedno značenje, što je veoma zahtjevan proces. Također nadodaje da kod interlingua modela postoji previše uvjeta za mnoge parove jezika. Zbog toga se razvio sustav temeljen na transferu, koji eksplicitno inzistira na pravilima transfera specifičnima za određene parove jezika.

Bhattacharyya (2015)[30] navodi dvije velike prepreke zbog kojih je povezivanje svih jezika u interlingua modelu nedostižan cilj. Prvo, jezici opisuju koncepte koristeći različite razine detalja. Drugo, pojmove sastavljene od više riječi teško je povezati s prirodnim leksemima koji opisuju istu pojavu na nekom drugom jeziku.

S druge strane, sustav temeljen na transferu proces analize može prilagoditi potrebnoj razini s obzirom na sličnost jezika koji su uključeni u prijevod. Bliski jezici zahtijevaju nisku razinu transfera, a različitiji nešto višu, te je za njih potrebno stvoriti mnoga različita pravila koja bi prijevod trebao slijediti pod različitim uvjetima.

[28] Poibeau, T. (2017). *Machine translation*. The MIT Press.
[29] Bhattacharyya, P. (2015). *Machine translation*. Chapman & Hall/CRC.
[30] Ibid.

3.3. Metoda transfera

Glavna značajka sustava temeljenih na metodi transfera je upotreba eksplicitnih pravila transfera. Ona se primjenjuju na paru jezika u procesu prijevoda, a njihova svrha je razrješavanje razlika u strukturi jezika, prema Bhattacharyya (2015)[31]. Arnold i sur. (1994)[32] opažaju da kod sustava temeljenih na metodi transfera postoji jasna potreba za često kompleksnim pravilima mapiranja među apstraktnim reprezentacijama izvornih i ciljnih rečenica.

Bhattacharyya (2015)[33] u svom radu opisuje tri glavna koraka u metodi transfera:

1. korak – analiza:

Tijekom analize, sustav analizira tekst na izvornom jeziku koristeći pritom pravila morfološke analize, parsiranja, generiranja semantike i mnoga druga. U ovom koraku bitno je istaknuti problem uklanjanja višeznačnosti, koja se pojavljuje u više oblika, kao što su višeznačnost lema, morfoloških struktura, općih i vlastitih imenica, leksička višeznačnost dijelova rečenice, te višeznačnost kod fraza i drugih izraza koji sadrže više riječi. Postoji i pragmatična višeznačnost namjere govornika i njegovog stava.

2. korak – transfer:

U ovom koraku sustav pretražuje riječi i fraze u dvojezičnom rječniku. Ako je korak analize bio uspješan u razrješavanju višeznačnosti u tekstu, te ako je dvojezični rječnik dovoljno detaljan i iscrpan, transfer će biti jednostavan. On podrazumijeva sam prijevod riječi i gramatičkih pravila. Međutim, nemoguće je eliminirati sve višeznačnosti u koraku analize. Također, dvojezična mapiranja ne mogu u potpunosti pokriti sve lekseme. Neka mapiranja su kompleksna, ili čak nemoguća, jer jezici prikazuju različite koncepte koristeći različite stupnjeve detalja. Takve specifičnosti u jezicima mogu dovesti do pogrešaka u ovom postupku.

[31] Bhattacharyya, P. (2015). *Machine translation*. Chapman & Hall/CRC.
[32] Arnold, D., Balkan, L., Meijer, S., Lee Humphreys, R., Sadler, L. (1994). *Machine translation: an introductory guide*. Blackwells NCC.

[33] Bhattacharyya, P. (2015). *Machine translation*. Chapman & Hall/CRC.

3. korak – generiranje:

Tijekom generiranja, sustav provodi morfološku sintezu lema koje pronalazi u dvojezičnim rječnicima, i zatim definira sintaktički poredak, tj. razmješta riječi i fraze na odgovarajuće pozicije u rečenici. Pritom se podrazumijeva poštivanje sintaktičkih pravila ciljnog jezika.

Prema Hutchinsu (1993)[34], pristupi strojnom prevođenju temeljeni na pravilima vode se pretpostavkom da prevođenje zahtijeva analizu i prikaz značenja teksta na izvornom jeziku i generaciju tog teksta na ciljnom jeziku. Ali iz primjera različitih sustava jasno je vidljivo da je za to potrebno puno više od prevođenja pojedinačnih riječi i definiranja pravila koja sustav za prevođenje treba poštivati.

[34] Hutchins, W.J. (1993). *Latest Developments in Machine Translation Technology: Beginning a New Era in MT Research.*

4. Statistički sustavi za strojno prevođenje

Budući da je razvoj interneta omogućio dostupnost velikih količina višejezičnog teksta, i zahvaljujući razvoju računala, polje strojnog prevođenja također se naglo razvijalo. Poibeau (2017)[35] uočava da je upravo to omogućilo razvoj statističkih sustava koji ovise o velikim količinama dvojezičnih tekstova. Raniji sustavi imali su direktan pristup prevođenju – tražili su ekvivalente riječi između dva jezika analizirajući podatke iz dvojezičnih rječnika. Danas se više ne baziraju na pojedinačnim riječima, već imaju sposobnost uočavanja nizova riječi, poput složenica, idioma, ili nizova nepovezanih riječi koje trebaju biti prevedene kao cjelina.

Hutchins (1995)[36] detaljnije opisuje statističke sustave za strojno prevođenje. Oni slijede principe statističkih metoda, koje su glavni temelj analize i generacije teksta. Ova metoda funkcionira tako što se u prvom koraku povezuju riječi, fraze i skupine riječi iz paralelnih prijevoda teksta, tj. dvojezičnih korpusa. Zatim sustav izračunava vjerojatnost koliko svaka riječ u rečenici izvornog jezika odgovara riječi ili skupini riječi u toj rečenici na ciljnom jeziku. Na temelju tih podataka generira se prijevod.

4.1. Važnost dvojezičnih korpusa

Poibeau (2017)[37] navodi dva pristupa sustavima za prevođenje na temelju korpusa: sustav temeljen na primjerima i statistički sustav. Sustav temeljen na primjerima temelji se na analizi postojećih prijevoda koji služe kao primjeri za nove prijevode. S druge strane, statistički pristup omogućuje izravnu izgradnju statističkih modela za strojno prevođenje koji se temelje na ogromnim količinama prevedenog teksta.

Koehn (2009)[38] ističe kako je jedna od najbitnijih značajki statističkih sustava za strojno prevođenje to što obično moraju biti trenirani na paralelnim korpusima. Paralelni korpus je skupina tekstova koji su upareni s prijevodom na ciljni jezik. Budući da je kompliciranо izgraditi sustav za prevođenje općenitih tekstova, mnogi sustavi za statističko strojno

[35] Poibeau, T. (2017). *Machine translation*. The MIT Press.
[36] Hutchins, W. J. (1995). Machine Translation: A Brief History. *Concise History of the Language Sciences*, 431–445.
[37] Ibid.
[38] Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

prevođenje razvijaju se za određenu domenu (Seljan, 2000)[39], pri čemu ograničavanje domene znatno pojednostavljuje postupak strojnog prevođenja zbog korištenja ograničenog vokabulara.

Tvorac Googleovog statističkog sustava za strojno prevođenje, Franz Josef Och, smatra da bi se dobra osnova za razvoj upotrebljivog sustava za novi jezični par trebala sastojati od dvojezičnog korpusa koji sadrži više od milijun riječi i dva jednojezična korpusa od po više od milijarde riječi (Brkić i sur., 2009)[40].

Statistički sustavi automatski prikupljaju specifična pravila iz dvojezičnih i jednojezičnih korpusa. Iako svi ti sustavi imaju isto temeljno načelo, razlikuju se u strukturi i izvorima svojih prijevodnih modela. Primjerice, postoji model temeljen na riječima, model temeljen na frazama, te modeli koji se temelje na sintaksi. Kod modela temeljenom na riječima, riječi se tretiraju kao leksemi, neovisno o ostalim riječima. Takav pristup morfologiji jedan je od glavnih nedostataka ovog modela, budući da sustav može prepoznati određeni oblik riječi, ali ne i drugačiji oblik iste riječi (Brkić i sur., 2009[41]).

Model temeljen na frazama sadrži rječnik temeljen na frazama (Och i Ney, 2004)[42] te prevodi kratke nizove riječi odjednom. Pritom se ne koriste sintaktičke ili morfološke informacije i pravila (Brkić i sur., 2009[43]).

Modeli temeljeni na sintaksi mogu se klasificirati prema njihovom sintaktičkom formalizmu, a najčešće korišteni model zasnovan je na strukturi sintaktičkog stabla. Prema Brkić i sur. (2009)[44], takvi modeli daju rezultate slične modelima temeljenim na frazama. Međutim, smatra se da bolje rezultate daje pristup koji kombinira statistički model i model temeljen na pravilima. (Brkić i sur., 2009[45]; Sepesy Maučec i Kačić, 2007[46]).

---

[39] Seljan, S. (2000). Sublanguage in Machine Translation. Mipro 2000: Computers in Intelligent Systems
[40] Brkić, M., Seljan, S., Vičić, T. (2009). Evaluation of the statistical machine translation service for Croatian-English. INFuture 2009: Digital resources and knowledge sharing.
[41] Ibid.
[42] Och, F., Ney, H. (2004). The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics, 30*, 417-449.
[43] Brkić, M., Seljan, S., Vičić, T. (2009). Evaluation of the statistical machine translation service for Croatian-English. INFuture 2009: Digital resources and knowledge sharing.
[44] Ibid.
[45] Ibid.
[46] Sepesy Maučec, M., Kačić, Z. (2007). Statistical Machine Translation from Slovenian to English. *Journal of computing and information technology, 15* (1), 47-59.

Sustavi za strojno prevođenje suočeni su s određenim preprekama u prikupljanju kvalitetnih korpusa. Prema Poibeau (2017)[47], postoje službeni izvori paralelnih korpusa, primjerice od strane institucija ili vlada država s više službenih jezika – one imaju javno dostupne pravne tekstove na svakom od svojih jezika. Primjerice, Europska Unija ima javno dostupne višejezične dokumente na temelju kojih je izgrađen Europarl korpus, koji sadrži više od 20 europskih jezika. No čak i kod takvih korpusa količina teksta varira ovisno o jezicima ili parovima jezika koji su zastupljeni. Nadalje, budući da su takvi tekstovi većinom fokusirani na domenu politike ili prava, sustavi za strojno prevođenje koji ih koriste najvjerojatnije neće uspješno prevoditi tekst iz drugih, nepovezanih domena, ili tekst nepovezane tematike.

Koehn (2009)[48] navodi da je upravo kontekst jedan od razloga zašto je teško izgraditi sustav za prijevod općenitih tekstova. Budući da riječi i fraze imaju različita značenja u različitim domenama, prijevod mora biti prilagođen stilu pisanja u domeni za koju se koristi.

Zato je potrebno stvarati nove korpuse, prilagođene domeni, posebno za manje zastupljene parove jezika.

Prema Koehnu (2009)[49], ako korisnik samostalno želi prikupiti paralelne korpuse koristeći primjerice tekstove prevedenih web stranica, može naići na mnoge probleme. Potrebno je odrediti koji dokumenti na prvom jeziku odgovaraju kojim dokumentima na drugom jeziku. To nije jednostavno jer direktno i potpuno prevedene web stranice nisu česta pojava. Primjerice, kod vijesti prevedenih na više jezika, tekst nije direktno preveden, već je prilagođen ciljanoj publici, izostavljajući tako neke dijelove originalnog teksta ili dodavajući nove.

Poibeau (2017)[50] ističe kako je automatska obrada teksta moguća samo ako imamo pristup ogromnim količinama podataka, tj. teksta. Zbog toga je među stručnjacima zastupljeno mišljenje da je najbolji pristup izgradnji statističkih sustava zapravo prikupljanje što više podataka. Oni moraju biti reprezentativni i raznovrsni, ali budući da je kvalitativne kriterije teško vrednovati, kvantitativni kriteriji još uvijek su puno bitnija stavka.

[47] Poibeau, T. (2017). *Machine translation*. The MIT Press.
[48] Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.
[49] Ibid.
[50] Poibeau, T. (2017). *Machine translation*. The MIT Press.

Iz toga slijedi da je dostupnost paralelnih korpusa jedna od najbitnijih stavki za izgradnju uspješnog sustava za statističko prevođenje. Za neke parove jezika postoji mnoštvo dostupnih korpusa, dok za neke, koji se rijetko koriste, ne postoji velika količina adekvatnih korpusa i zato su se razvile tehnike automatskog prikupljanja korpusa, najčešće prikupljanjem teksta s interneta. Međutim, mnogi jezici nisu dovoljno zastupljeni na internetu, a još manje su dostupni njihovi izravni prijevodi na neki drugi jezik, uočava Poibeau (2017)[51].

Budući da kvaliteta paralelnih korpusa izravno utječe na kvalitetu automatskog strojnog prijevoda (Seljan i Dunđer, 2015a[52]), tj. na performanse sustava za strojno prevođenje, očita je potreba za izgradnjom kvalitetnih i opsežnih paralelnih korpusa. Načelno vrijedi, što su podatkovni skupovi koji se koriste pri izgradnji sustava za strojno prevođenje kvalitetniji i reprezentativniji, to će i strojni prijevodi u konačnici biti točniji i precizniji (Dunđer, 2015)[53].

4.2. Poravnavanje rečenica

Dunđer (2015)[54] zapaža da se rečenično sravnjivanje korpusa može izvršiti ručno ili automatski pomoću raznih metoda, primjenom postojećih online besplatnih ili komercijalnih alata (Seljan i sur., 2008[55]; Brkić i sur., 2011[56]; Seljan i sur., 2010[57]).

Prema Poibeau (2017)[58], pretpostavka je da prijevod slijedi strukturu originalnog teksta i da su rečenice najčešće formirane na isti način u izvornom i ciljnom jeziku. Također, u većini slučajeva postoji korelacija duljine rečenica na izvornom i ciljnom jeziku. Prema tome, automatski je moguće poravnati rečenice na temelju relativne razlike u duljini rečenica. Kako bi se izbjegle pogreške u poravnavanju, potrebno je sagledati tekst u cjelini, ne samo rečenicu po rečenicu.

[51] Poibeau, T. (2017). *Machine translation*. The MIT Press.
[52] Seljan, S., Dunđer, I. (2015a). Automatic quality evaluation of machine-translated output in sociological-philosophical-spiritual domain. Central European Conference on Information and Intelligent Systems (CECIIS).
[53] Dunđer, I. (2015). *Sustav za statističko strojno prevođenje i računalna adaptacija domene.* Doktorska disertacija. Sveučilište u Zagrebu. Filozofski fakultet.
[54] Ibid.
[55] Seljan, S., Agić, Ž., Tadić, M. (2008). Evaluating sentence alignment on Croatian-English parallel corpora. Proceedings of the 6th International Conference on Formal Approaches to South Slavic and Balkan Languages.
[56] Brkić, M., Matetić, M., Seljan, S. (2011). Towards Obtaining High Quality Sentence-Aligned English-Croatian Parallel Corpus. Proceedings of the 4th IEEE International Conference on Computer Science and Information Technology ICCSIT 2011.
[57] Seljan, S., Tadić, M., Agić, Ž., Šnajder, J., Bašić, B.D., Osmann, V. (2010). Corpus Aligner (CorAl) Evaluation on English-Croatian Parallel Corpora. *LREC*.
[58] Poibeau, T. (2017). *Machine translation*. The MIT Press.

Poibeau (2017)[59] dalje navodi da je moguće pronaći slične riječi u izvornom i ciljnom tekstu, koje služe kao točke povezanosti (engl. *correspondence points*). Najčešće su to riječi poput osobnih imena, lokacija, vlastitih imenica, brojeva i akronima. Parovi rečenica sa više točki povezanosti su najvjerojatnije međusobni prijevodi, i tako računalo može automatski eliminirati neusklađenosti u poravnavanju. Ovakav pristup naziva se leksičkim pristupom jer se temelji na analizi dijela leksika, tj. rječnika.

U idealnom slučaju rečenice u korpusima trebale bi se podudarati što je više moguće. Jedna rečenica u izvornom jeziku trebala bi odgovarati jednoj rečenici u ciljnom jeziku, zaključuje Poibeau (2017)[60]. Najefikasniji pristup je zapravo kombiniranje leksičkog pristupa sa pristupom temeljenim na duljini rečenica (Seljan i sur., 2010[61]). Samo duljina rečenica, kao ni samo točke povezanosti, nisu dovoljne kako bi se poravnala dva teksta.

Osim poravnavanja rečenica postoji i problem poravnavanja samih riječi unutar rečenice. Taj zadatak je teško automatizirati jer jedna riječ iz izvornog jezika može biti povezana s mnogim riječima ili skupinama riječi na ciljnom jeziku. Statistički sustavi analiziraju dvojezične korpuse koji sadrže više prijevoda za jednu riječ u drugačijim kontekstima. Svaki prijevod dobiva ocjenu vjerojatnosti da je upravo to točan prijevod. Zbog toga paralelni korpusi čine temelj statističkih sustava Poibeau (2017)[62].

4.3. Treniranje sustava

Kako bi treniranje bilo uspješno, Poibeau (2017)[63] dodaje da su osim paralelnih korpusa potrebni i dvojezični rječnici. Treniranjem, sustavi se poboljšavaju, a ovaj pristup koristi se i u strojnom i u strojno potpomognutom prevođenju (Brkić i sur., 2009[64]). Oni se sastoje od riječi na izvornom jeziku i svih mogućih prijevoda te riječi na ciljnom jeziku. Svakom prijevodu izračunava se ocjena na temelju vjerojatnosti da je upravo ta riječ točan prijevod u određenom kontekstu. Faza treniranja ili učenja još se naziva i fazom kodiranja (engl. *encoding phase*), budući da uključuje kodiranje informacija o jeziku.

[59] Poibeau, T. (2017). *Machine translation*. The MIT Press.
[60] Ibid.
[61] Seljan, S., Tadić, M., Agić, Ž., Šnajder, J., Bašić, B.D., Osmann, V. (2010). Corpus Aligner (CorAl) Evaluation on English-Croatian Parallel Corpora. *LREC*.
[62] Poibeau, T. (2017). *Machine translation*. The MIT Press.
[63] Ibid.
[64] Brkić, M., Seljan, S., Vičić, T. (2009). Evaluation of the statistical machine translation service for Croatian-English. INFuture 2009: Digital resources and knowledge sharing.

Ključne komponente statističkog sustava za prevođenje su jezični model (engl. *language model*) i prijevodni model (engl. *translation model*).

Dunđer (2015)[65] opisuje jezični model kao funkciju koja uzima rečenicu na ciljnom jeziku i daje vjerojatnost da je ta rečenica u „duhu jezika“, tj. da odgovara stvarnoj upotrebi. Drugim riječima, jezični model procjenjuje koliko je vjerojatan jedan segment rečenice u ciljnom jeziku.

Prema Koehnu (2009)[66], jezični model utječe na izbor ispravnih riječi, ispravan poredak riječi u rečenici i druge odluke sustava. On dodjeljuje svakoj rečenici rezultat koliko je vjerojatno da bi takvu rečenicu izgovorio izvorni govornik, tj. ocjenjuje tečnost rečenice.

Dunđer (2015)[67] dalje objašnjava kako prijevodni model uparuje riječi, odnosno nizove riječi, iz izvornog jezika u ciljni jezik, tj. procjenjuje podudarnost leksičkih jedinica u izvornom i ciljnom jeziku. Stoga se razvija pomoću dvojezičnog sravnjenog paralelnog korpusa koji se sastoji od segmenata na izvornom jeziku te ljudski prevedenih prijevoda na ciljnom jeziku. Prijevodni model opisuje vjerojatnost prijevoda na temelju paralelnog korpusa, s obzirom na to da bez korpusa nije moguće izravno procijeniti vjerojatnost prijevodnih parova. Načelno vrijedi, što je veća vjerojatnost, to je izglednije da se radi o dobrom prijevodu.

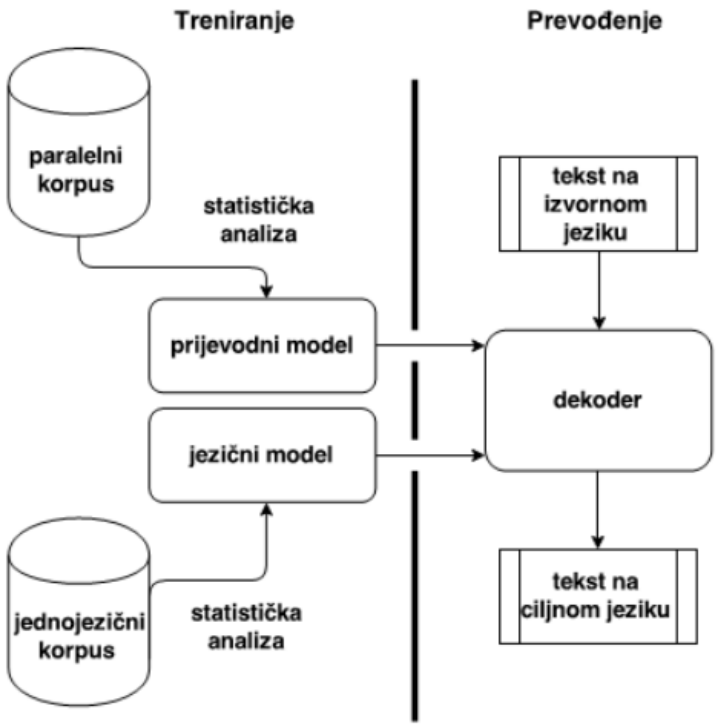
4.4. Testiranje sustava

U fazi testiranja koriste se informacije iz dvojezičnog korpusa kako bi se prevele nove rečenice. Poibeau (2017)[68] navodi da se ova faza također se naziva fazom dekodiranja (engl. *decoder phase*) jer sustav nastoji dekodirati ulaznu rečenicu. Svaku rečenicu na izvornom jeziku sustav rastavi na riječi, pretražuje najvjerojatniji prijevod svake riječi, uzimajući u obzir i poredak riječi u ciljnom jeziku.

Dunđer (2015)[69] u svom radu opisuje princip rada dekodera. Statistički modeli svakom mogućem prijevodu pridružuju određene vrijednosti, tj. vjerojatnosti (eng. *score*) s obzirom na dani niz riječi u izvornom jeziku. Vjerojatnosti prevođenja fraze i preslagivanja, tj.

[65] Dunđer, I. (2015). *Sustav za statističko strojno prevođenje i računalna adaptacija domene.* Doktorska disertacija. Sveučilište u Zagrebu. Filozofski fakultet.
[66] Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.
[67] Dunđer, I. (2015). *Sustav za statističko strojno prevođenje i računalna adaptacija domene.* Doktorska disertacija. Sveučilište u Zagrebu. Filozofski fakultet.
[68] Poibeau, T. (2017). *Machine translation*. The MIT Press.
[69] Dunđer, I. (2015). *Sustav za statističko strojno prevođenje i računalna adaptacija domene.* Doktorska disertacija. Sveučilište u Zagrebu. Filozofski fakultet.

premještanja redoslijeda riječi te vjerojatnost jezičnog modela se objedinjuju kako bi se svakom mogućem prijevodu pridružio jedan združeni rezultat, tj. konačna vjerojatnost (eng. *final score*). Dekoder zatim treba pronaći najbolji prijevod iz skupa mogućih prijevoda. Može se reći da je zadaća dekodera u modelu statističkog strojnog prevođenja za izvornu rečenicu pronaći prijevod u ciljnom jeziku s maksimalnom vjerojatnošću.



Slika 2. Model statističkog strojnog prevođenja (Dunđer, 2015)[70]

Statistički sustavi za strojno prevođenje predstavljaju značajan napredak u odnosu na dotadašnje sustave. Razvijali su se zajedno s tehnološkim naprecima, a njihovom razvoju posebno je pridonijela dostupnost velike količine višejezičnih tekstova. Oni čine temelj sustava za strojno prevođenje, a njihov utjecaj je i danas značajan za suvremene sustave.

[70] Dunđer, I. (2015). *Sustav za statističko strojno prevođenje i računalna adaptacija domene.* Doktorska disertacija. Sveučilište u Zagrebu. Filozofski fakultet.

4.5. Evaluacija sustava

Automatska evaluacija sustava za strojno prevođenje predmet je mnogih istraživanja, osobito kada su u pitanju manje zastupljeni, morfološki bogati jezici (Brkić i sur., 2013)[71]. Pri automatskoj evaluaciji koriste se razne metrike. Pomoću njih se nastoji vrednovati kvaliteta prijevoda slijedeći standarde ljudskog vrednovanja (Seljan i Dunđer, 2015b)[72].

Prema Seljan i sur. (2015)[73], kod online sustava za strojno prevođenje kvaliteta prijevoda ovisi o parovima jezika koji se koriste i o vrsti teksta koji se prevodi. Nadalje, postoji i problem zastupljenosti jezika - određeni parovi jezika manje su zastupljeni jer za njih nije prikupljeno dovoljno paralelnih korpusa. Kvaliteta prevedenih tekstova također ovisi i o sličnosti jezika.

Seljan i Dunđer (2015)[74] u svom su istraživanju koristili automatske metrike BLEU, NIST, METEOR i GTM za evaluaciju javno dostupnih sustava za strojno prevođenje, Google Translate i Yandex.Translate. Provedena je evaluacija prijevoda između dvaju jezičnih parova: englesko-hrvatskog i rusko-hrvatskog. Budući da su ruski i hrvatski blisko povezani slavenski jezici, istraživanje je pokazalo bolje rezultate za rusko-hrvatski jezični par za oba alata. Dobiveni rezultati slični su prosječnim rezultatima ljudske evaluacije.

Rezultate automatskih metrika također je moguće poboljšati prethodnom obradom, koja uključuje prijepis teksta u mala slova i tokenizaciju (Brkić i sur., 2013)[75].

U drugom istraživanju Seljan i sur. (2015)[76], prikazani su rezultati ljudske evaluacije strojno prevedenih tekstova za dva jezična para, englesko-hrvatski i rusko-hrvatski. Ljudska evaluacija provedena je pomoću kriterija fluentnosti i adekvatnosti. Fluentnost označava koliko je kvalitetno tekst preveden s obzirom na oblikovanje, stil i lingvističke standarde ciljnog jezika, dok adekvatnost označava koliko je dobro tekst preveden s obzirom na očuvanje značenja izvorne rečenice.

[71] Brkić, M., Seljan, S., Vičić, T. (2013). Automatic and human evaluation on English-Croatian legislative test set. Intelligent Text Processing and Computational Linguistics, Springer.
[72] Seljan, S., Dunđer, I. (2015b). Machine Translation and Automatic Evaluation of English/Russian-Croatian. Proceedings of the International Conference "Corpus Linguistics". St. Petersburg State University, 72-79.
[73] Seljan, S., Tucaković, M., Dunđer, I. (2015). Human Evaluation of Online Machine Translation Services for English/Russian-Croatian. *New Contributions in Information Systems and Technologies.*
[74] Seljan, S., Dunđer, I. (2015b). Machine Translation and Automatic Evaluation of English/Russian-Croatian. Proceedings of the International Conference "Corpus Linguistics". St. Petersburg State University, 72-79.
[75] Brkić, M., Seljan, S., Vičić, T. (2013). Automatic and human evaluation on English-Croatian legislative test set. Intelligent Text Processing and Computational Linguistics, Springer.
[76] Seljan, S., Tucaković, M., Dunđer, I. (2015). Human Evaluation of Online Machine Translation Services for English/Russian-Croatian. *New Contributions in Information Systems and Technologies.*

Kod oba jezična para, prevedeni tekstovi postigli su bolje rezultate s obzirom na kriterij adekvatnosti nego kod kriterija fluentnosti. Rezultati upućuju na to da oba sustava češće nailaze na probleme s formiranjem rečenica i poretkom riječi nego sa semantikom, tj. značenjem. Nadalje, utvrđeno je da su najčešći tip pogrešaka morfološke pogreške. Sintaktičke pogreške predstavljale su problem u prijevodima s engleskog, dok su neprevedene ili izostavljene riječi i leksičke pogreške bile češće u prijevodima s ruskog na hrvatski jezik (Seljan i sur., 2015[77]).

[77] Seljan, S., Tucaković, M., Dunđer, I. (2015). Human Evaluation of Online Machine Translation Services for English/Russian-Croatian. *New Contributions in Information Systems and Technologies.*

5. Neuronski sustavi za strojno prevođenje

Iako su sustavi za strojno prevođenje znatno napredovali od svojih početaka, oni se još uvijek susreću s određenim problemima.

Prema Koehnu (2020)[78], računala, baš poput ljudi, u procesu prijevoda uzimaju u obzir kontekst rečenice, tj. značenje cjelokupne rečenice, a ne samo pojedinačnih riječi. Značenje nije uvijek doslovno. Primjerice, idiomatski izrazi ne mogu se prevoditi doslovno i pritom sačuvati originalno značenje.

Također postoje i sintaktičke prepreke, primjerice kod prevođenja jezika koji imaju znatno drugačiju sintaktičku strukturu. No Koehn (2020)[79] navodi da semantika predstavlja još veću prepreku u prevođenju. To dolazi do izražaja primjerice u rečenicama gdje je značenje samo implicirano. Ljudski prevoditelji svjesni su takvih posebnosti jezika jer posjeduju znanje o svijetu, kontekstu i pravilima jezika na koji prevode.

Sustavi za strojno prevođenje također zahtijevaju neku razinu takvog znanja. Stoga je razvojem računala i sustava za strojno prevođenje došlo do početka primjene umjetne inteligencije u procesu strojnog prevođenja, s ciljem da dobiveni prijevodi budu što točniji i fluentniji, baš poput ljudskih prijevoda.

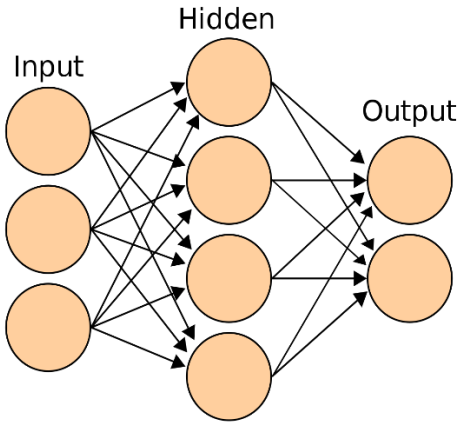
5.1. Duboko učenje kod sustava za strojno prevođenje

Duboko učenje (engl. *deep learning*) koristi se u mnoge svrhe, kao što su prepoznavanje slika, obrada govora, te obrada prirodnog jezika, navodi Poibeau (2017)[80]. Iako se te namjene međusobno razlikuju, sam proces dubokog učenja obično funkcionira tako da sustav iz velikog broja primjera automatski izdvaja najrelevantnije karakteristike, koje se kod dubokog učenja nazivaju značajke (engl. *features*). Značajka je apstraktno svojstvo koje neuronska mreža automatski uočava analizom uzoraka koji se ponavljaju unutar korpusa.

[78] Koehn, P. (2020). *Neural machine translation*. Cambridge University Press.
[79] Ibid.
[80] Poibeau, T. (2017). *Machine translation*. The MIT Press.

Poibeau (2017)[81] također ističe da se učenje odvija na principu hijerarhije – počinje se od osnovnih elemenata, primjerice znakova ili riječi kad govorimo o jezicima, i na temelju njih se određuju kompleksnije strukture, poput nizova riječi ili fraza. Nakon toga se dolazi do cjelokupne analize objekta u pitanju, primjerice cijele rečenice. Taj proces je sličan procesima koji se odvijaju u ljudskom mozgu.

Neuronski sustavi za strojno prevođenje temelje se na umjetnim neuronskim mrežama. Prema Koehnu (2020)[82], umjetne neuronske mreže (engl. *artificial neural networks*) inspirirane su neuronima u ljudskom mozgu. Ukratko, neuronska mreža na temelju određenog broja ulaznih vrijednosti (engl. *input*) predviđa izlaznu vrijednost (engl. *output*).



Slika 3. Primjer umjetne neuronske mreže (Wikipedia, 2006)[83]

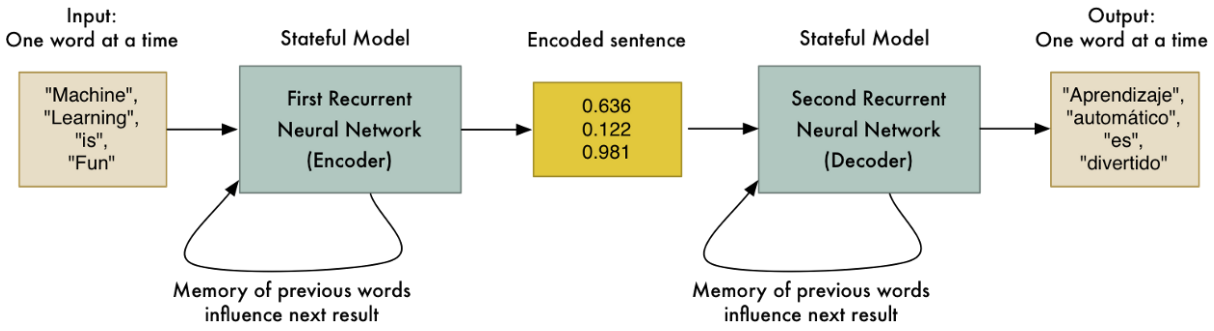
Kod strojnog prevođenja, Poibeau (2017)[84] navodi da je prednost dubokog učenja mogućnost izgradnje sustava u kojima se mali broj elemenata treba ručno specificirati. Cilj je izgraditi sustav koji samostalno uči, pronalazi najbolja rješenja i donosi zaključke na temelju analize podataka. Primjerice, sustav bi trebao automatski uočiti sintaktičke odnose u rečenici iako ne kodira sintaksu izravno. Statistički sustavi za prevođenje također su se vodili tom idejom, ali u stvarnosti se mnogi parametri moraju ručno definirati i prilagođavati kako bi sustav bio uspješan.

[81] Poibeau, T. (2017). *Machine translation*. The MIT Press.
[82] Koehn, P. (2020). *Neural machine translation*. Cambridge University Press.
[83] En.wikipedia.org. (2006). *Artificial neural network*. Dostupno na: https://en.wikipedia.org/wiki/Artificial_neural_network
[84] Poibeau, T. (2017). *Machine translation*. The MIT Press.

Prema Poibeau (2017)[85], kod dubokog učenja, računalo može donositi pouzdane odluke zahvaljujući mogućnosti analiziranja raznih vrsta informacija istovremeno. Ovakvi modeli su hijerarhijski, ali mogu se opisati i kao višedimenzionalni – svaki element, poput riječi ili fraze, osim što ima svoje mjesto u hijerarhiji, smješten je i unutar šireg konteksta. Nadalje, Poibeau (2017)[86] navodi da su neuronski sustavi kontinuirani. Oni proučavaju cijelu rečenicu bez da ju je potrebno rastaviti na manje dijelove kao kod statističkih sustava. Na temelju te ideje, neuronski sustavi mogu uspoređivati riječi, ali i fraze i rečenice, te uočavati povezanosti između njih.

5.2. Struktura neuronskih modela

Cho i sur. (2014)[87] prikazali su strukturu neuronskih modela. Neuronski modeli za strojno prevođenje najčešće se sastoje od kodera (engl. *encoder*) i dekodera (engl. *decoder*). Koder obrađuje ulaznu rečenicu (engl. *input sentence*) promjenjive duljine i na temelju te rečenice stvara njezinu vektorsku reprezentaciju fiksne duljine. Prema toj reprezentaciji dekoder generira izlaznu rečenicu (engl. *output sentence*), tj. točan prijevod promjenjive duljine na ciljnom jeziku. Dekoder je moguće trenirati za predviđanje sljedeće riječi s obzirom na kontekstni vektor i sve prethodno predviđene riječi.



Slika 4. Primjer arhitekture neuronskog modela (Revanuru i sur., 2017)[88]

[85] Poibeau, T. (2017). *Machine translation*. The MIT Press.
[86] Ibid.
[87] Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*.
[88] Revanuru, K., Turlapty, K., Rao, S. (2017). Neural Machine Translation of Indian Languages. IIIT Bangalore, ACM Compute.

5.3. Treniranje modela

Modeli se mogu trenirati na dvojezičnim i jednojezičnim korpusima, te dvojezičnim rječnicima. Yang i sur. (2020)[89] analiziraju taj proces i navode da se prije unosa podataka za treniranje u model, riječi trebaju kodirati u vektore, tj. oblik koji neuronska mreža može obrađivati. Tijekom treniranja, u model se unose korpusi, s ciljem da sustav može uspješno mapirati ulazne rečenice s odgovarajućim izlaznim rečenicama na ciljnom jeziku. Ulazna rečenica kodira se u svoju vektorsku reprezentaciju, a neuronska mreža obrađuje podskup podataka koji sadrži uzorke za treniranje.

5.4. Generiranje prijevoda

Prema Yang i sur. (2020)[90], nakon treniranja modela, on se može koristiti za generiranje novih prijevoda (engl. *inference*). Proces generiranja sličan je procesu treniranja, s razlikom da kod generiranja postoji tzv. skriveno stanje kodera (engl. *encoder hidden state*) kojemu se ne može pristupiti.

Tan i sur. (2020)[91] nadovezuju se i pobliže opisuju generiranje prijevoda. Neuronski modeli za strojno prevođenje najčešće koriste arhitekturu kodera i dekodera. Nju čine četiri glavne komponente: slojevi reprezentacije (engl. *embedding layers*), mreže kodera i dekodera (engl. *coder and decoder networks*), te klasifikacijski sloj (engl. *classification layer*).

Sloj reprezentacije temelji se na konceptu kontinuirane reprezentacije (engl. *continuous representation*). On mapira odvojene simbole u jedan kontinuirani vektor.

Prema Cho i sur. (2014)[92], koder obrađuje ulaznu rečenicu promjenjive duljine i na temelju te rečenice stvara njezinu vektorsku reprezentaciju fiksne duljine. Prema toj reprezentaciji, dekoder generira točan prijevod promjenjive duljine na ciljnom jeziku.

Poibeau (2017)[93] dalje navodi da je svaka riječ kodirana pomoću vektora, a sve vektorske reprezentacije tih riječi se na kraju kombiniraju i čine reprezentaciju cijele rečenice.

[89] Yang, S., Wang, Y., Chu, X. (2020). A Survey of Deep Learning Techniques for Neural Machine Translation.
[90] Ibid.
[91] Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. *AI Open*, *1*, 5–21.
[92] Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*.
[93] Poibeau, T. (2017). *Machine translation*. The MIT Press.

Prema Tan i sur. (2020)[94], klasifikacijski sloj predviđa distribuciju pojavnica (engl. *token*) na ciljnom jeziku.

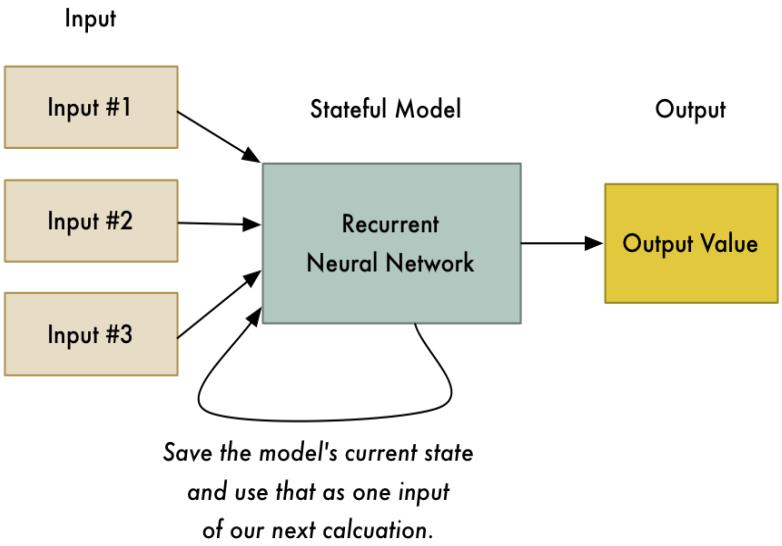
Yang i sur. (2020)[95] zaključuju da se taj proces naziva prijevodom od početka do kraja (engl. *end-to-end translation*) jer struktura koder-dekoder prevodi tekst iz izvornih podataka izravno u ciljni jezik, što znači da dio tog procesa, tj. skriveni sloj, nije vidljiv.

Postoje mnoge metode izgradnje arhitekture kodera i dekodera, a prema Tan i sur. (2020)[96], mogu se podijeliti u tri glavne kategorije. To su:

1. Metode temeljene na rekurentnim neuronskim mrežama (engl. *recurrent neural network - RNN*)
2. Metode temeljene na konvolucijskim neuronskim mrežama (engl. *convolutional neural network - CNN*)
3. Metode temeljene na mrežama samopozornosti (engl. *self-attention network - SAN*)

Prema Tan i sur. (2020)[97], rekurentne neuronske mreže jedne su od najsnažnijih oblika neuronskih mreža.

[94] Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. *AI Open*, *1*, 5–21.
[95] Yang, S., Wang, Y., Chu, X. (2020). A Survey of Deep Learning Techniques for Neural Machine Translation.
[96] Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. *AI Open*, *1*, 5–21.
[97] Ibid.



Slika 5. Primjer arhitekture rekurentne neuronske mreže (Revanuru i sur., 2017)[98]

Yang i sur. (2020)[99] dalje analiziraju rekurentne neuronske mreže. Ističu kako one mogu imati više slojeva. Mreže s jednim slojem obično daju lošije rezultate prijevoda nego one s više slojeva.

Navode da se rekurentne neuronske mreže mogu razlikovati i prema usmjerenosti. Postoje jednosmjerne i dvosmjerne mreže. Dvosmjerna mreža često poboljšava kvalitetu prijevoda. One se koriste u jednom sloju mreže kako bi zabilježile informacije o kontekstu, budući da model daje kvalitetniji prijevod ako su mu dostupne informacije o kontekstu. U takvom sustavu, prvi sloj čita rečenicu s lijeva na desno, a drugi sloj čita istu rečenicu s desna na lijevo. Zatim te rečenice služe kao informacije za idući sloj.

Yang i sur. (2020)[100] također navode da je vjerojatnost pojavljivanja neke riječi određena ostalim riječima koje se pojavljuju prije ili poslije nje. Kod upotrebe jednosmjernih mreža, teško je zabilježiti ovisnost prve riječi o zadnjoj, jer vektori prolaze kroz prevelik broj stanja. Ali dvosmjerne mreže imaju dodatni sloj informacija s obrnutim slijedom čitanja riječi, što smanjuje relativnu udaljenost među riječima.

[98] Revanuru, K., Turlapty, K., Rao, S. (2017). Neural Machine Translation of Indian Languages. IIIT Bangalore, ACM Compute.
[99] Yang, S., Wang, Y., Chu, X. (2020). A Survey of Deep Learning Techniques for Neural Machine Translation.
[100] Ibid.

Tan i sur. (2020)[101] zaključuju da se neuronski sustavi za strojno prevođenje znatno razlikuju od prethodnih pristupa strojnom prevođenju. Oni koriste jednu veliku neuronsku mrežu pri modeliranju cijelog procesa prevođenja. Zbog toga nema potrebe za određivanjem velikog broja pojedinačnih značajki sustava. Nadalje, ističu da ovakvi sustavi koriste kontinuirane prikaze umjesto zasebnih simboličkih prikaza koje koriste statistički sustavi. Također, treniranje neuronskih sustava temelji se na pristupu od početka do kraja (engl. *end-to-end*), za razliku od statističkih sustava kod kojih svaka komponenta treba biti zasebno podešena. Neuronski sustavi za strojno prevođenje postigli su vrhunske rezultate na primjerima mnogih jezičnih parova.

Iako i kod dubokog učenja postoje pogreške u prijevodima, ono omogućava nove načine rješavanja određenih problema s kojima se susreću sustavi za strojno prevođenje i zato je doživjelo veliki uspjeh u domeni prevođenja.

5.5. Evaluacija sustava

Budući da sustavi za strojno prevođenje ne daju savršene rezultate, postoji potreba za evaluacijom sustava i kvalitete prijevoda. Kvaliteta prijevoda je subjektivna, ali ona obično predstavlja sličnost, tj. podudaranje između strojnog prijevoda i referentnog, tj. ljudskog prijevoda (Dunđer i sur., 2020)[102].

Seljan i sur. (2020)[103] ističu da je kvaliteta prijevoda književnog teksta od velike važnosti za istraživače u području visokog obrazovanja, osobito one koji se bave proučavanjem jezika i književnosti. Strojne prijevode zbog toga je potrebno vrednovati i najčešće naknadno urediti, tj. ispraviti u skladu s potrebama korisnika.

U istraživanju Seljan i sur. (2020)[104] provedena je ljudska evaluacija strojnih prijevoda poezije za hrvatsko-njemački jezični par. Evaluacija je provedena koristeći kriterije adekvatnosti i

[101] Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. *AI Open*, *1*, 5–21.
[102] Dunđer, I., Seljan, S., Pavlovski, M. (2020). Automatic Machine Translation of Poetry and a Low-Resource Language Pair. MIPRO 2020.
[103] Seljan, S., Dunđer, I., Pavlovski, M. (2020). Human Quality Evaluation of Machine-Translated Poetry. MIPRO 2020.
[104] Ibid.

fluentnosti. Istraživanje Martindale i sur. (2019)[105] pokazalo je da neuronski sustavi za strojno prevođenje postižu dobre rezultate prema kriteriju fluentnosti, ali prevedeni tekstovi pritom ne zadovoljavaju uvijek i kriterij adekvatnosti. Seljan i sur. (2020)[106] u svom radu dobivaju drugačije rezultate – prevedeni tekstovi imaju bolje rezultate prema kriteriju adekvatnosti nego prema kriteriju fluentnosti.

Iako korišteni neuronski sustavi nisu posebno izgrađeni za domenu poezije, pokazalo se da prijevodi poezije s njemačkog na hrvatski zadovoljavaju kriterije adekvatnosti i fluentnosti. Međutim, u većini slučajeva prijevodi bi trebali biti naknadno uređeni kako bi se kvaliteta dodatno poboljšala.

Dunđer i sur. (2020)[107] u svom su radu koristili automatske metrike za evaluaciju prijevoda poezije za njemačko-hrvatski jezični par. Njihovi rezultati pokazali su da korišteni sustavi za strojno prevođenje ne daju dovoljno dobre rezultate kod prijevoda u domeni poezije, što je vidljivo u dobivenim ocjenama metrika BLEU, METEOR, RIBES i CharacTER.

Kvaliteta prijevoda ključna je za razumijevanje izvornog teksta, posebno kod kompliciranih, dvosmislenih i idiomatskih tekstova kao što su književna djela. Kako bi se postigla bolja kvaliteta prijevoda, potrebno je provesti ljudsku evaluaciju kvalitete strojno prevedenog teksta, uzimajući u obzir kriterije adekvatnosti i fluentnosti (Seljan i sur., 2020)[108].

[105] Martindale, M., Carpuat, M., Duh, K., McNamee, P. (2019). Identifying Fluently Inadequate Output in Neural and Statistical Machine Translation, Proc. of Machine Translation Summit XVII volume 1: Research Track.
[106] Seljan, S., Dunđer, I., Pavlovski, M. (2020). Human Quality Evaluation of Machine-Translated Poetry. MIPRO 2020.
[107] Dunđer, I., Seljan, S., Pavlovski, M. (2020). Automatic Machine Translation of Poetry and a Low-Resource Language Pair. MIPRO 2020.
[108] Seljan, S., Dunđer, I., Pavlovski, M. (2020). Human Quality Evaluation of Machine-Translated Poetry. MIPRO 2020.

6. Istraživanje

U praktičnom dijelu rada opisan je postupak izgradnje vlastitog sustava za strojno prevođenje u odabranoj domeni, za hrvatsko-engleski jezični par. Provedeno je prikupljanje i priprema podataka potrebnih za treniranje i za izgradnju sustava. Na kraju je provedena automatska i ručna evaluacija dobivenih prijevoda.

6.1. Prikupljanje podataka za treniranje

Kako bi se pomoću ovog alata izgradio vlastiti sustav za strojno prevođenje, prvo je potrebno prikupiti podatke za treniranje sustava. To su prvenstveno dvojezični korpusi, a moguće je koristiti i prijevodne memorije, kao i dvojezične rječnike. Budući da ovakvi sustavi daju kvalitetnije prijevode ako su izgrađeni za određenu domenu, prikupljeni su materijali samo iz specijalizirane domene zdravstva, za hrvatsko-engleski jezični par. U svrhu izgradnje ovog sustava korišteno je ukupno 10 289 parova rečenica. Od toga je 5116 samostalno prikupljenih rečenica, preuzetih s javno dostupnih web stranica u Hrvatskoj i inozemstvu. Preostale 5173 rečenice preuzete su u sklopu gotovih prijevodnih memorija dostupnih za javno korištenje.

6.2. Izgradnja sustava

Sustav za strojno prevođenje zahtijeva nekoliko vrsta podataka: dokumente za treniranje (engl. *training*), ugađanje (engl. *tuning*) i testiranje (engl. *testing*) sustava.

U dokumente za treniranje uključeni su prvenstveno dvojezični korpusi, prijevodne memorije i dvojezični rječnici.

Dokumente za ugađanje čini reprezentativni uzorak sadržaja koji korisnik namjerava prevoditi koristeći sustav.

Dokumenti za testiranje su dvojezični korpusi u kojima je tekst na ciljnom jeziku najtočniji prijevod odgovarajućih rečenica na izvornom jeziku. Oni ne utječu na kvalitetu izgrađenog sustava za prevođenje, već na temelju njih sustav određuje rezultat BLEU metrike.

BLEU metrika, kao jedna od standardnih automatskih metrika, prikazuje koliko se automatski prijevod koji je generirao izgrađeni sustav podudara s referentnim prijevodima iz dokumenata za testiranje. BLEU rezultat koji je na kraju dodijeljen izgrađenom sustavu je prosječni BLEU rezultat svih rečenica iz dokumenata za testiranje. Što je viši BLEU rezultat, to je podudarnost automatskog prijevoda sa referentnim viša. Ako je model treniran na jednoj usko definiranoj domeni, BLEU rezultat bit će viši nego kod sustava izgrađenih za općenite prijevode. Dokumente za testiranje korisnik može postaviti sam, ili sustav može odabrati dio rečenica iz dvojezičnih korpusa za treniranje.

Prema Brkić i sur. (2012)[109], BLEU metrika funkcionira tako što podudara n-grame dobivenog prijevoda s n-gramima referentnog prijevoda i broji podudaranja na razini rečenice. Te rečenice se zbrajaju unutar čitavog seta rečenica za testiranje. Podudaranja ne ovise o položaju u rečenici. Adekvatnost se računa u preciznosti riječi, dok se tečnost računa u preciznosti n-grama. Konačni rezultat BLEU metrike je geometrijski prosjek modificiranih točnosti n-grama. BLEU rezultati kreću se od 0 do 100, odnosno od 0 do 1.

Prema Denkowski i Lavie (2010)[110], BLEU rezultati iznad 30 bodova općenito odražavaju razumljive prijevode, a BLEU rezultati iznad 50 bodova odražavaju dobre i tečne prijevode. BLEU metrika, budući da je statistički utemeljena i neovisna o jeziku, ne uzima u obzir morfološke varijante riječi. Ova metrika zahtijeva točno podudaranje riječi, pri čemu su sva podudaranja jednako ocijenjena (Brkić i sur., 2012)[111].

Međutim, Brkić i sur. (2011)[112] i Koehn (2010)[113] ističu i određene kritike usmjerene na ovu metriku. Najbitnije je napomenuti da ona zanemaruje relativnu važnost riječi te ne uzima u obzir ukupnu gramatičku koherentnost. Ljudski BLEU rezultati jedva su viši od rezultata sustava za strojno prevođenje iako su ljudski prijevodi mnogo bolje kvalitete. Nadalje, BLEU

[109] Brkić, M., Seljan, S., Vičić, T. (2012). BLEU Evaluation of Machine-Translated EnglishCroatian Legislation. LREC, 2143-2148.
[110] Denkowski, M., Lavie, A. (2010). Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks. *Proceedings of the Association for Machine Translation in the Americas AMTA*.
[111] Brkić, M., Seljan, S., Vičić, T. (2012). BLEU Evaluation of Machine-Translated EnglishCroatian Legislation. LREC, 2143-2148.
[112] Brkić, M., Seljan, S., Matetić, M. (2011). Machine translation evaluation for croatian-english and English-Croatian language pairs. NLPCS Workshop: Human-Machine Interaction in Translation. Copenhagen: Copenhagen Business School.
[113] Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

metrika nije intuitivna i oslanja se na veliki broj rečenica kako bi bila u korelaciji s ljudskim evaluacijama (Brkić i sur., 2011[114]; Snover i sur., 2006[115]).

Budući da su upravo dokumenti korišteni za treniranje sustava reprezentativni uzorak sadržaja kojemu je ovaj sustav namijenjen, u ovom slučaju sustav samostalno odabire rečenice za ugađanje i testiranje.

Idući korak je izgradnja modela. Nakon toga započinje proces treniranja. Nakon što je treniranje uspješno završeno, očitava se rezultat BLEU metrike. Moguće je i ručno usporediti prijevode koje je napravio izgrađeni model s referentnim prijevodima iz podataka za testiranje.

[114] Brkić, M., Seljan, S., Matetić, M. (2011). Machine translation evaluation for Croatian-English and English-Croatian language pairs. NLPCS Workshop: Human-Machine Interaction in Translation. Copenhagen: Copenhagen Business School.
[115] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of Association for Machine Translation in the Americas. AMTA.

6.3. Rezultati

Language Pair: Croatian - English
Category: Healthcare

Data | Models

Name	Status	Modified Date	Bleu Score	Baseline Bleu Score	Training	Dictionary	Tuning	Test
model_zdravstvo	Training succeeded	2021-06-22	▲ 50.49	45.06	9,145	0	506	481

Slika 6. Rezultati evaluacije sustava

Nakon uspješne izgradnje sustava, ovaj model postigao je BLEU rezultat koji iznosi 50,49 bodova. Taj rezultat uspoređen je s polaznim BLEU rezultatom (engl. *baseline BLEU score*), koji iznosi 45,06 bodova. Polazni BLEU rezultat je rezultat prijevoda iste skupine rečenica za testiranje koje je preveo alat Microsoft Bing Translator.

Izgrađeni model postigao je rezultat bolji za 05,43 boda jer je treniran na određenoj domeni, čime je vokabular korišten za treniranje ograničen i specijaliziran.

6.4. Vrednovanje kvalitete prijevoda

U ovom koraku odabrano je 10 rečenica iz podataka za testiranje, te su prijevodi koje je generirao izgrađeni sustav uspoređeni s referentnim prijevodima i s prijevodima online alata Microsoft Translator. Većina prijevoda bila je točna i uspjela je prenijeti originalno značenje, iako dobivene rečenice nisu bile potpuno iste kao referentni prijevod.

Tablica 1. Primjer prijevoda 1

Istraživanja su pokazala da djeca koja se koriste znakovima češće razgovaraju s roditeljima.	*Referentni prijevod:* Research has shown that children who use signs talk to their parents more often.
	Vlastiti sustav: Research has shown that children who use the signs are more likely to talk to parents.
	Microsoft alat: Studies have shown that children using signs are more likely to talk to parents.

Na prvom primjeru vidljivo je da je prijevod vlastitog sustava sličan referentnom prijevodu. Jedina razlika je u prijevodu riječi „češće“, za koju referentni prijevod glasi „more often“, dok prijevod vlastitog sustava i Microsoft alata glasi „more likely“. Značenje nije identično, ali smisao rečenice je očuvan i prijevod je tečan.

Tablica 2. Primjer prijevoda 2

Svi možemo doprinijeti zaustavljanju širenja COVID-19.	*Referentni prijevod:* We can all do our bit to help stop the spread of COVID-19.
	Vlastiti sustav: Everyone can help stop the spread of COVID-19.
	Microsoft alat: We can all contribute to stopping the spread of COVID-19.

U drugom primjeru također postoje razlike između referentnog i vlastitog prijevoda. Najuočljivija razlika je kod riječi „doprinijeti“. Referentni prijevod u ovom slučaju prati svakodnevni stil govora, što je vidljivo u sintagmi „do our bit“ (učiniti svoje). Vlastiti sustav prevodi riječ „doprinijeti“ u „help“, što je također točan prijevod. Naposlijetku, Microsoft prevoditelj koristi riječ „contribute“, što bi zapravo bio najtočniji prijevod riječi „doprinijeti“. Tu su vidljive razlike u stilu rečenice, ali svaka verzija prijevoda je točna i prenosi istu poruku.

Tablica 3. Primjer prijevoda 3

Nuspojave na koje sumnjate da su nastale od cjepiva možete prijaviti davatelju cjepiva.	*Referentni prijevod:* You can report suspected side effects to your vaccination provider.
	Vlastiti sustav: Suspected side effects can be reported to your vaccine provider.
	Microsoft alat: You can report side effects that you suspect are from vaccines to your vaccine provider.

U trećem primjeru vidljiva je razlika u gramatici – rečenica na izvornom jeziku koristi glagol u aktivnom obliku „možete prijaviti“, kao i referentni prijevod „you can report“. S druge strane, vlastiti sustav koristi pasivni oblik „can be reported“. Uzrok tome najvjerojatnije su podaci iz dvojezičnih korpusa koji većinom sadrže službene informacije, a samim time koriste i službeni stil, kod kojeg je pasivni oblik glagola česta pojava.

Tablica 4. Primjer prijevoda 4

Caklina gubi bjelinu i postaje sivkasta ili žućkasta.	*Referentni prijevod:* The enamel loses its whiteness and becomes grayish or yellowish.
	Vlastiti sustav: The enamel loses whiteness and becomes grayish or yellowish.
	Microsoft alat: Enamel loses whiteness and becomes grayish or yellowish.

Četvrti primjer je prijevod kratke rečenice koju je vlastiti sustav preveo gotovo identično referentnom prijevodu, uz jednu razliku: kod izraza „gubi bjelinu“, referentni prijevod dodaje i zamjenicu „svoju“ – “loses its whiteness”, dok vlastiti prijevod izostavlja zamjenicu, što daje rezultat „loses whiteness“. Zamjenica u ovom slučaju zapravo nije potrebna jer je subjekt vidljiv u samoj rečenici, te se podrazumijeva da se izostavljena zamjenica “svoju” odnosi na subjekt „caklina“.

Tablica 5. Primjer prijevoda 5

Tijekom farmakološke terapije uobičajene su nuspojave; polovica pacijenata osjeća simptome slične gripi, a kod jedne trećine bilježe se emocionalni problemi.	*Referentni prijevod:* Adverse effects with these treatments were common, with half of people getting flu-like symptoms and a third experiencing emotional problems.
	Vlastiti sustav: Side effects are common during pharmacological therapy; half of patients experience flu-like symptoms, and one-third experience emotional problems.
	Microsoft alat: Side effects are common during pharmacological therapy; half of patients experience flu-like symptoms, and one-third experience emotional problems.

Rečenica iz petog primjera dulja je od ostalih, što rezultira većim brojem razlika između referentnog i vlastitog prijevoda. Pojam „nuspojave“ referentni prijevod navodi kao „adverse effects“, dok vlastiti prijevod i Microsoft prevoditelj koriste pojam „side effects“. Oba prijevoda zapravo imaju isto značenje. Također, referentni prijevod izostavlja pojam „farmakološke terapije“. Moguće je da se on već pojavljuje u prethodnoj rečenici u samom tekstu pa je značenje jasno iz konteksta. No vlastiti sustav svejedno prevodi taj pojam u „pharmacological therapy“.

Tablica 6. Primjer prijevoda 6

Vodite računa o oznakama na tlu.	*Referentni prijevod:* Pay attention to floor markers.
	Vlastiti sustav: Take care of the markings on the ground.
	Microsoft alat: Take care of the markings on the ground.

U šestom primjeru izgrađeni sustav napravio je pogrešku. Frazu „vodite računa“ prevodi kao „take care“. To u ovom određenom kontekstu nije točan prijevod, koji bi trebao glasiti „pay attention“, kao u referentnom prijevodu. U ovom je primjeru problematična semantika jer obje fraze imaju slično značenje, ali u ovom kontekstu samo je referentni prijevod točan.

Tablica 7. Primjer prijevoda 7

Moguće je da će ubuduće dom morati na neko vrijeme još više ograničiti broj osoba koje dolaze u posjete domu.	*Referentni prijevod:* There may come a time when your aged care facility needs to restrict people visiting your home further.
	Vlastiti sustav: It is possible that in the future the home may need to further limit the number of people visiting the home for a while.
	Microsoft alat: It is possible that in the future the home will have to limit even more the number of people visiting the home for a while.

Sedmi prijevod sadrži isti problem sa semantikom. Riječ „dom“ u ovom kontekstu se odnosi na staračk dom, što je u referentnom prijevodu prevedeno kao „aged care facility“. Vlastiti sustav doslovno je preveo riječ dom u „home“, što nije točan prijevod u ovom kontekstu.

Tablica 8. Primjer prijevoda 8

I kada je umro, rekli su mi da je umro od začepljenja vena.	*Referentni prijevod:* And when he died, they told me it was a thrombosis.
	Vlastiti sustav: And when he died, they told me that he died from a blockage of his veins.
	Microsoft alat: And when he died, they told me he died of a blockage of veins.

U osmom primjeru također je jedna riječ drugačije prevedena. Pojam „začepljenje vena“ u referentnom prijevodu prevedeno je kao „thrombosis“, što je stručan termin za to stanje. Ali vlastiti sustav, kao i Microsoft prevoditelj, prevode taj pojam kao „blockage of veins“, što je doslovan prijevod. Značenje se iz njega može jasno iščitati, ali budući da je sustav namijenjen prevođenju u domeni zdravstva, trebao bi moći prevoditi i stručne termine.

Tablica 9. Primjer prijevoda 9

Pretraga se ne preporučuje kod pacijenata mlađih od tri godine i kod pacijenata s kontrakturama stopala.	*Referentni prijevod:* The test is not recommended for patients younger than three years or for patients with foot contractures.
	Vlastiti sustav: The examination is not recommended in patients younger than three years of age and in patients with foot contractions.
	Microsoft alat: The test is not recommended in patients under three years of age and in patients with foot contractures.

U devetom primjeru prijevod vlastitog sustava je točan, iako postoji razlika u jednoj riječi. Referentni prijevod pojam „pretraga“ navodi kao „test“, a vlastiti sustav navodi taj pojam kao „examination“. Unatoč ovoj razlici u izboru riječi, oba prijevoda imaju isto značenje.

Tablica 10. Primjer prijevoda 10

No, prekomjerna i neprimjerena uporaba antibiotika ubrzava pojavljivanje i širenje bakterija rezistentnih na antibiotike.	*Referentni prijevod:* However, excessive and inappropriate use of antibiotics accelerates the emergence and spread of antibiotic-resistant bacteria.
	Vlastiti sustav: However, the over-use and inappropriate use of antibiotics accelerates the emergence and spread of antibiotic-resistant bacteria.
	Microsoft alat: But the excessive and inappropriate use of antibiotics accelerates the appearance and spread of antibiotic-resistant bacteria.

U desetom primjeru prikazan je prijevod rečenice koju je vlastiti sustav također preveo potpuno točno, iako postoji jedna razlika. Prijevod pojma „prekomjerna uporaba“ u referentnom prijevodu glasi „excessive use“, dok je u prijevodu vlastitog sustava taj pojam preveden kao „over-use“. Unatoč ovoj razlici, oba prijevoda su točna.

7. Zaključak

Cilj ovoga rada bio je prikazati različite pristupe strojnom prevođenju te provesti evaluaciju vlastitog izrađenog sustava za strojno prevođenje.

Rezultati su pokazali da izgrađeni sustav u većini slučajeva uspješno prevodi rečenice iz odabrane domene, s prosječnom ocjenom BLEU= 50,49. Dobiveni prijevodi često se razlikuju od referentnih prijevoda. Primjerice, razlike se javljaju kod odabira pojedinih riječi, a često i kod redoslijeda riječi u rečenici. Međutim, većina prijevoda je točna i fluentna bez obzira na te razlike, a neki prijevodi su identični referentnom prijevodu. Prijevodi koji nisu bili uspješni često sadrže semantičke pogreške. Primjerice, pojedine riječi bile su doslovno prevedene, što nije bio ispravni prijevod u određenom kontekstu u kojemu se pojavljuju. Također, uočene su pogreške kod prijevoda specifičnih riječi koje ne pripadaju odabranoj domeni. Budući da korpusi korišteni za treniranje sustava većinom sadrže vokabular iz jedne domene, možemo zaključiti da sustav u mnogim situacijama ne može uspješno prevoditi tekst iz drugih domena.

Kod vlastitog izgrađenog sustava većina prijevoda bila je točna, a pogreške koje su se javljale nisu previše utjecale na razumljivost prijevoda. Prijevodi koje generiraju profesionalni sustavi za strojno prevođenje kvalitetniji su, iako ni oni ne daju uvijek savršene rezultate. Kroz povijest su doživjeli su značajan napredak, što je dovelo do korištenja online alata za strojno prevođenje u svakodnevnoj komunikaciji.

U teorijskom dijelu rada prikazana su tri najznačajnija pristupa sustavima za strojno prevođenje: sustav temeljen na pravilima, statistički sustav i neuronski sustav za strojno prevođenje. Opisane su najvažnije karakteristike svakog pristupa i načela na kojima se temelje. Zatim je u praktičnom dijelu prikazan proces izgradnje vlastitog sustava iz određene domene, pomoću odabranog alata. Na kraju je provedena evaluacija sustava pomoću BLEU metrike, kao i ljudska evaluacija.

8. Literatura

1. Arnold, D., Balkan, L., Meijer, S., Lee Humphreys, R., Sadler, L. (1994). *Machine translation: an introductory guide*. Blackwells NCC.
2. Bhattacharyya, P. (2015). *Machine translation*. Chapman & Hall/CRC.
3. Brkić, M., Matetić, M., Seljan, S. (2011). Towards Obtaining High Quality Sentence-Aligned English-Croatian Parallel Corpus. Proceedings of the 4th IEEE International Conference on Computer Science and Information Technology ICCSIT 2011.
4. Brkić, M., Seljan, S., Matetić, M. (2011). Machine translation evaluation for Croatian-English and English-Croatian language pairs. NLPCS Workshop: Human-Machine Interaction in Translation. Copenhagen: Copenhagen Business School.
5. Brkić, M., Seljan, S., Vičić, T. (2009). Evaluation of the statistical machine translation service for Croatian-English. INFuture 2009: Digital resources and knowledge sharing.
6. Brkić, M., Seljan, S., Vičić, T. (2012). BLEU Evaluation of Machine-Translated English-Croatian Legislation. LREC, 2143-2148.
7. Brkić, M., Seljan, S., Vičić, T. (2013). Automatic and human evaluation on English-Croatian legislative test set. Intelligent Text Processing and Computational Linguistics, Springer.
8. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. https://doi.org/10.3115/v1/w14-4012 (Pristupljeno: 29. lipanj 2021.)
9. Denkowski, M., Lavie, A. (2010). Choosing the Right Evaluation for Machine Translation: An Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks. *Proceedings of the Association for Machine Translation in the Americas AMTA.*
10. Dovedan, Z., Seljan, S., Vučković, K. (2002). Strojno prevođenje kao pomoć u procesu komunikacije. *Informatologia*, 35(4).
11. Dunđer, I. (2015). *Sustav za statističko strojno prevođenje i računalna adaptacija domene.* Doktorska disertacija. Sveučilište u Zagrebu. Filozofski fakultet.

12. Dunđer, I., Seljan, S., Arambašić, M. (2013). Domain-specific Evaluation of Croatian Speech Synthesis in CALL. Recent Advances in Information Science - Computer Engineering, WSEAS 1, 142-147.
13. Dunđer, I., Seljan, S., Pavlovski, M. (2020). Automatic Machine Translation of Poetry and a Low-Resource Language Pair. MIPRO 2020.
14. En.wikipedia.org. (2006). *Artificial neural network*. https://en.wikipedia.org/wiki/Artificial_neural_network (Pristupljeno: 26. kolovoz 2021.)
15. Hutchins, W.J. (1993). *Latest Developments in Machine Translation Technology: Beginning a New Era in MT Research.*
16. Hutchins, W. J. (1995). Machine Translation: A Brief History. *Concise History of the Language Sciences*, 431–445. https://doi.org/10.1016/b978-0-08-042580-1.50066-0 (Pristupljeno: 29. lipanj 2021.)
17. Koehn, P. (2020). *Neural machine translation*. Cambridge University Press.
18. Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.
19. Kučiš, V., Seljan, S. (2014). The role of online translation tools in language education. *Babel*, *60*(3). https://doi.org/10.1075/babel.60.3.03kuc (Pristupljeno: 26. kolovoz 2021.)
20. Martindale, M., Carpuat, M., Duh, K., McNamee, P. (2019). Identifying Fluently Inadequate Output in Neural and Statistical Machine Translation, Proc. of Machine Translation Summit XVII volume 1: Research Track.
21. Och, F., Ney, H. (2004). The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics, 30*, 417-449.
22. Poibeau, T. (2017). *Machine translation*. The MIT Press.
23. Revanuru, K., Turlapty, K., Rao, S. (2017). Neural Machine Translation of Indian Languages. IIIT Bangalore, ACM Compute.
24. Seljan, S. (2000). Sublanguage in Machine Translation. Mipro 2000: Computers in Intelligent Systems.
25. Seljan, S. (2011). Translation technology as Challenge in education and business. *Informatologia*, *44*(4), 279–286.
26. Seljan, S. (2018a). Total Quality Management Practice in Croatian Language Service Provider Companies. EntreNova 18, 4 (1), 461-469. INFuture2015: eInstitutions–Openness, Accessibility, and Preservation.

27. Seljan, S. (2018b). Quality Assurance (QA) of Terminology in a Translation Quality Management System (QMS) in the Business Environment. European Parliament: Translation Services in the Digital World, 92-145.
28. Seljan, S. (2019) Informacijska i komunikacijska tehnologija (IKT) u interdisciplinarnom okruženju nastave jezika. U: Vrhovac, Y., Berlengi Kapučin, V., Geld, R., Jelić, A., Letica Krevelj, S., Mardečić, S., Lütze-Miculinić, M. (ur.) *Izazovi učenja stranoga jezika u osnovnoj školi* (pp. 446–461). Ljevak.
29. Seljan, S., Agić, Ž., Tadić, M. (2008). Evaluating sentence alignment on Croatian-English parallel corpora. Proceedings of the 6th International Conference on Formal Approaches to South Slavic and Balkan Languages.
30. Seljan, S., Dunđer, I. (2014). Combined Automatic Speech Recognition and Machine Translation in Business Correspondence Domain for English-Croatian. *International Journal of Computer and Information Engineering, World Academy of Science, Engineering and Technology, 8*(11)*,* 1980-1986.
31. Seljan, S., Dunđer, I. (2015a). Automatic quality evaluation of machine-translated output in sociological-philosophical-spiritual domain. Central European Conference on Information and Intelligent Systems (CECIIS).
32. Seljan, S., Dunđer, I. (2015b). Machine Translation and Automatic Evaluation of English/Russian-Croatian. Proceedings of the International Conference "Corpus Linguistics". St. Petersburg State University, 72-79.
33. Seljan, S., Dunđer, I., Pavlovski, M. (2020). Human Quality Evaluation of Machine-Translated Poetry. MIPRO 2020.
34. Seljan, S., Katalinić, J. (2017). Integrating Localization into a Video Game. INFuture2017: Integrating ICT in Society.
35. Seljan, S., Klasnić, K., Stojanac, M., Pešorda, B., Mikelić Preradović, N. (2015). Information Transfer through Online Summarizing and Translation Technology. INFuture2015: e-Institutions–Openness, Accessibility, and Preservation.
36. Seljan, S., Tucaković, M., Dunđer, I. (2015). Human Evaluation of Online Machine Translation Services for English/Russian-Croatian. *New Contributions in Information Systems and Technologies.*
37. Seljan, S., Škof Erdelja, N., Kučiš, V., Dunđer, I., Pejić Bach, M. (2020). Quality Assurance in Computer-Assisted Translation in Business Environments. Natural Language Processing for Global and Local Business. IGI-Global, 242-270.

38. Seljan, S., Tadić, M., Agić, Ž., Šnajder, J., Bašić, B.D., Osmann, V. (2010). Corpus Aligner (CorAl) Evaluation on English-Croatian Parallel Corpora. *LREC*.
39. Sepesy Maučec, M., Kačić, Z. (2007). Statistical Machine Translation from Slovenian to English. *Journal of computing and information technology, 15* (1), 47-59. https://doi.org/10.2498/cit.1000760 (Pristupljeno: 26. kolovoz 2021.)
40. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of Association for Machine Translation in the Americas. AMTA.
41. Sreelekha, S., Bhattacharyya, P., Malathi, D. (2017). Statistical vs. Rule-Based Machine Translation: A Comparative Study on Indian Languages. *Advances in Intelligent Systems and Computing*, 663–675. https://doi.org/10.1007/978-981-10-5520-1_59 (Pristupljeno: 29. lipanj 2021.)
42. Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. *AI Open*, *1*, 5–21. https://doi.org/10.1016/j.aiopen.2020.11.001 (Pristupljeno: 29. lipanj 2021.)
43. Tripathi, S. i Sarkhel, J. (2011). Approaches to machine translation. *Annals of Library and Information Studies, 57.* 388-393.
44. Yang, S., Wang, Y., Chu, X. (2020). A Survey of Deep Learning Techniques for Neural Machine Translation. *ArXiv, abs/2002.07526*. (Pristupljeno: 29. lipanj 2021.)

9. Popis slika

10. Popis tablica

Strojno prevođenje u odabranoj domeni

Sažetak

Cilj ovog rada je prikaz različitih pristupa sustavima za strojno prevođenje, te izgradnja vlastitog sustava za strojno prevođenje iz odabrane domene, za hrvatsko-engleski jezični par. Rad je sastavljen od dva osnovna dijela, teorijskog i praktičnog. U teorijskom dijelu prikazane su različite arhitekture sustava za strojno prevođenje: sustava temeljenog na pravilima, statističkog sustava i neuronskog sustava. U drugom, praktičnom dijelu, analiziran je proces izgradnje vlastitog sustava za strojno prevođenje koji uključuje prikupljanje resursa, izgradnju sustava i zatim evaluaciju te vrednovanje kvalitete dobivenog strojnog prijevoda.

Ključne riječi: strojno prevođenje, izgradnja sustava, domena, evaluacija

Machine translation in a selected domain

Summary

The aim of this thesis is to present different approaches to machine translation systems, as well as to construct a custom machine translation system in a selected domain, for the Croatian-English language pair. In the theoretical part of the thesis, different types of architectures of machine translation systems are presented: the rule-based system, the statistical system, and the neural system. In the practical part, the process of building a custom machine translation system is analyzed. It includes gathering necessary resources and the construction of the system, followed by evaluation of the quality of the obtained machine translations.

Key words: machine translation, custom system construction, domain, evaluation